Introduction

The empirical-apportion phenomenon of solubility of water into the partition coefficient (LogP) is widely used as a measure of lipophilicity in the assessment of the movement of phases in physical or biological systems of new entities. In the case of agrochemicals LogP serves a valuable role in the evaluation of their environmental risk. Wide establishment in the chemical industry as a useful and frequently considered property gave rise to a great number of available methods and software applications suitable for the prediction of octanol-water partition coefficient for new compounds. In this study we have compared six generally acknowledged structure based methods for the prediction of LogP:

- CLogP (Daylight v4.73)
- AlogP (Acosymmetry Diamond Descriptors v1.5)
- LogP (Phys Chem Batch v8.16)
- Kowen (EPI Suite v1.12)
- ABLogP (Pharma Algorithms ADME Boxes v3.5)
- ABLogP (Pharma Algorithms ADME Boxes v3.5)

A comparison is also made with the Syngenta internal ELogP methodology which derives a consensus value from ACDLogP and AlogP rather than using single method for the evaluation of LogP. Finally a completely new algorithm developed by Pharma Algorithms is introduced – ABLogP 2.0 based on ‘Trainable Model’ methodology (available in Pharma Algorithms ADME Boxes v4.0).

Data Set & Methodology

Selected methods have been applied to a test set of 1000 compounds randomly selected from Syngenta research projects. Nearly 4000 additional compounds with measured LogP values from the same source were utilized in the investigation using ‘Trainable LogP’ method. Predictions for each obtained for the test set were compared to the Syngenta measured values in terms of R-squared, intercept, slope and mean absolute error values (MAE) calculated from plots normalized to give a slope of 1 and intercept of 0. In addition percentage average error values within 0.2, 0.5, 1.0 and 2.0 log units were assessed for each method.

Method Comparison

In the table the results are presented for the six listed structure based methods of LogP prediction along with new ABLogP 2.0 and the Syngenta consensus ELogP methods. The bar plot displays percentage of compounds in the 1000 molecule test set having predicted absolute errors within each of the selected threshold values.

As can be seen from the data presented above all of the selected structure based methods yield broadly comparable results on our test set. AlogP gives a somewhat weaker performance in terms of the percentage of the compounds having predicted absolute errors of up to 1 log unit but the difference from the other methods is really quite marginal. ALogP methodology performed noticeably better in all aspects than any of the individual methods it based on supporting the rationale of deriving a consensus value from a number of methods rather than trusting any single one of them. However it is still restricted by some of the limitations of the structure based methods it is derived from and from the obtained results it appears unlikely that any prediction method trained on literature data sets could give rise to an RI > 0.5 and MAE < 0.5 for novel ‘in house’ data sets.

Trainable LogP

The last method described in the Method Comparison discussion actually addresses one of the fundamental problems preventing the effective use of third-party prediction algorithms in the chemical industry, i.e. the literature based training set rarely covers the specific part of the chemical space occupied by the compounds that a certain company is working with or sometimes a specific experimental protocol used to measure the property of interest yields results contrasting with experimental values for the same compounds found in the training set. This is one of the reasons why a method should allow a company to effectively take a third-party predictive algorithm into its specific results using proprietary ‘in house’ data.

Addressing this issue Pharma Algorithms has developed a concept of ‘Trainable Models’ that provides a novel solution to this problem. Each ‘Trainable Model’ consists of the following parts:

- A structure based QSAR/QSPR for the prediction of a certain property derived from a literature training set for Pharma Algorithms – the so called baseline QSAR/QSPR
- Any user defined data set with the experimental values for the property of interest – the so called Self-training Library
- A series of ‘Trainable LogP’ models using different Self-training Libraries were derived in the manner described above and applied to the test set of 1000 Syngenta compounds:

<table>
<thead>
<tr>
<th>Method</th>
<th>R2</th>
<th>Slope</th>
<th>Intercept</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using Built-in Library</td>
<td>0.80</td>
<td>0.80</td>
<td>0.96</td>
<td>0.58</td>
</tr>
<tr>
<td>Built-in Library + 1000 Syngenta compounds</td>
<td>0.65</td>
<td>0.83</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>Built-in Library + 2500 Syngenta compounds</td>
<td>0.70</td>
<td>0.84</td>
<td>0.41</td>
<td>0.49</td>
</tr>
<tr>
<td>Built-in Library + c.a. 4000 Syngenta compounds</td>
<td>0.72</td>
<td>0.85</td>
<td>0.40</td>
<td>0.46</td>
</tr>
</tbody>
</table>

These presented results clearly show that ‘Trainable LogP’ model successfully copes with the task of training itself on specific compounds from the Syngenta database as the addition of growing numbers of such compounds to the Self-training Library of the model gave a steady increase in the accuracy of the predictions and improvement of the distribution of compounds according to the absolute error values.

As it can be seen the Reliability index (RI) closely correlates with the accuracy of the predictions and with the changes in the distribution according to the absolute error values. This fact justifies the use of the Reliability index as a measure of the model applicability domain. Moreover it can be noted that the enlarging of the Self-training Library gives not only the effect of rising accuracy but most importantly the increase of the share of better quality predictions or in other words the expansion of the models applicability domain. This is clearly demonstrated by the number of acceptable and high quality predictions shown in the parenthesis table and the histograms of compound number distribution according to the RI value for four ‘Trainable LogP’ models with growing libraries of increasing size presented here.

Acknowledgements

Authors wish to thank Ann Stanforth and Tom Shefki (Industrial Placement Students, University of Bath) for assistance with data analysis.

References